A:\WscV&V1.wpd
Printed on October 26, 1999 (3:59PM)
Written on July 8, 1999

VALIDATION OF MODELS: STATISTICAL TECHNIQUES AND DATA AVAILABILITY

Jack P.C. Kleijnen

Department of Information Systems (BIK)/Center for Economic Research (CentER)
School of Economics and Management (FEW)
Tilburg University (KUB)
Postbox 90153, 5000 LE Tilburg, The Netherlands.

Phone: +3113-4662029; fax: +3113-4663377; e-mail: kleijnen@kub.nl;
web: http://center.kub.nl/staff/kleijnen

Prepared as Reference No. MA07 for the
Simulation Model Validation session of the
Analysis Methodology Track of the
1999 Winter Simulation Conference

Pointe Hilton Resort, Squaw Peak, Phoenix, Arizona

December 5-8, 1999

VALIDATION OF MODELS: STATISTICAL TECHNIQUES AND DATA AVAILABILITY

Jack P.C. Kleijnen

Department of Information Systems/Center for Economic Research (CentER)
School of Economics and Management (FEW)
Tilburg University
5000 LE Tilburg, THE NETHERLANDS

ABSTRACT

This paper shows which statistical techniques can be used to validate simulation models, depending on which real-life data are available. Concerning this availability, three situations are distinguished (i) no data, (ii) only output data, and (iii) both input and output data. In case (i) - no real data - the analysts can still experiment with the simulation model to obtain simulated data; such an experiment should be guided by the statistical theory on the design of experiments. In case (ii) - only output data - real and simulated output data can be compared through the well-known two-sample Student t statistic or certain other statistics. In case (iii) - input and output data - trace-driven simulation becomes possible, but validation should not proceed in the popular way (make a scatter plot with real and simulated outputs, fit a line, and test whether that line has unit slope and passes through the origin); alternative regression and bootstrap procedures are presented. Several case studies are summarized, to illustrate the three types of situations.

1. INTRODUCTION

This paper gives a survey on how to validate simulation models through the application of *statistical* techniques, such that the type of technique actually applied depends on the *availability of data on the real system*. Regarding this data availability, I distinguish three situations:
(i) no real-life data are available,
(ii) there is only data on the real output (not the corresponding input or scenario),
(iii) besides the output data, the corresponding input or trace is also known, which is used to perform so-called trace driven or correlated inspection simulation (see Law and Kelton 1991, p. 316).

What, however, does 'validation' mean? A whole book could be written on the philosophical and practical issues involved in validation (see, for example, the monograph by Knepell, and Arangno 1993)! For this survey, however, I define *validation* as determining whether the simulation model is an acceptable representation of the real system - given the purpose of the simulation model (again see Law and Kelton 1991).

The literature on validation is abundant: see the web (http://manta.cs.vt.edu/biblio/), and the detailed surveys in Beck et al. (1997), Kleijnen (1995b), and Sargent (1996). In that literature, however, the focus in not on the role of data availability in the choice of statistical tests! This contribution has such a focus; it is a revision of Kleijnen (1999).

So I concentrate on validation that uses *mathematical statistics*. After all, simulation means experimentation (albeit with a model instead of the real system), and any experimentation calls for statistical analysis, preceded by statistical design. Obviously, such a statistical analysis is only part of the whole validation process (other parts are graphical summaries, animation for 'face validity', etc.; many types of validation are used and proposed in practice and theory; see the references at the end of this contribution). However, if mathematical statistics is used, then the correct statistics should be used!

Which type of statistical procedure is correct obviously depends on the kind of data that are available for analysis. Briefly, my *main conclusions* will turn out to be as follows.

Case (i): Even if real data are missing, there is still *expert knowledge*. (For example, we all are experts in waiting at supermarkets, so we know that if more customers arrive per hour, then waiting times increase - unless more cashiers become active.) However, this knowledge is *qualitative*; to obtain quantitative knowledge, a simulation model is developed (i.e., the sign or direction of the effect is known, not its magnitude). If the simulation model's input/output (I/O) behavior violates this qualitative knowledge, the model should be seriously questioned: are there programming and conceptual errors? In §2 I shall present a systematic method for selecting conditions or scenarios as input for the simulation model, namely, design of experiments or DOE. In practice; simulation errors have indeed been detected in this way.

Case (ii):If data on the real output are available,

then we can apply the classical two-sample Student t statistic - provided the data are approximately normally distributed. In case of non-normality we can use distribution-free tests or bootstrapping. See §3.

Case (iii): In trace-driven simulation we can apply a particular kind of regression analysis (compute the differences and sums of real and simulated outputs; regress these differences on the sums, and test for zero intercept and zero slope). In case of non-normality, however, bootstrapping of the difference between the average simulated and real outputs gives best results (prespecified type I error probability α and high power). See §4.

2. NO REAL DATA AVAILABLE: DOE

How realistic is it to assume that there is no data on the real system being simulated? Indeed, in some applications, such data are either completely missing or scarce. Examples are data on nuclear war (fortunately, no data, except for outdated figures on Hiroshima and Nagasaki), nuclear accidents (limited data: Chernobyl, Three Miles Island), global warming or greenhouse effect (few data; see Kleijnen, Van Ham, and Rotmans 1992, and Jansen and De Vries 1998).

If no data on the real system are available, then strong validation claims are impossible. Yet the analysts should at least perform sensitivity analysis (or what-if analysis). I define *sensitivity analysis* as the systematic investigation of the reaction of the simulation responses to *extreme* values of the model's input or to *drastic* changes in the model's structure. For example, what happens to the customers' mean waiting time when their arrival rate doubles; what happens if the priority rule is changed by introducing 'fast lanes'? (The literature does not provide a standard definition of sensitivity analysis; some authors consider only marginal changes of continuous inputs.)

I use the DOE term *factor* to denote a parameter, an input variable, or a module of a simulation model. In the supermarket example, a parameter is the arrival or service rate; an input variable is the number of cashiers; a module may be the submodel for the priority rules (First-In-First-Out or FIFO, priority for customers with less than - say - ten items).

Sensitivity analysis can support validation: such an analysis shows whether factors have effects that agree with experts' prior qualitative knowledge (for example, faster service gives lower mean waiting time). Admittedly, in practice not all simulation models have effects with known signs; yet, many models do have factors with known signs (as the case studies below will demonstrate).

Sensitivity analysis further shows which factors are important. If possible, information on these factors should be collected, for validation purposes (availability of such data enables trace-driven simulation; see §4). If the significant factors are controllable by the users, then sensitivity analysis shows how to change these factors to optimize the real system (see Kleijnen and Pala 1999 for an application).

The importance of sensitivity analysis in validation is also emphasized by Fossett et al. (1991), who present three military case studies, and Nayani and Mollaghasemi (1998), who present a semiconductor case study.

Sensitivity analysis of a simulation model requires a set of simulation runs. By definition, during a simulation run, all factors remain constant; simulated time increases, and in a stochastic simulation model a stream of pseudorandom numbers is generated. Factors do change from run to run; that is, each factor has at least two levels or 'values' in the experiment as a whole. The factor may be *qualitative*, as the priority rules exemplified. A detailed discussion of qualitative factors and various measurement scales is given in Kleijnen (1987, pp. 138-142).

There are several techniques for sensitivity analysis, Most practitioners change *one factor at a time*, and think that this is *the* scientific way to perform what-if analysis. Actually it is easy to prove mathematically that - compared with DOE's resolution-3 designs - this method gives less accurate estimates of a factor's first-order effect (called 'main effect' in ANOVA, Analysis Of Variance). Moreover, changing one factor at a time does not enable estimation of 'interactions' among factors: what happens if two or more factors change simultaneously? DOE's resolution-4 and resolution-5 designs enable the estimation of two-factor interactions, as we shall see next (the remainder of this section is based on Kleijnen 1998).

DOE's central problem is how to select a limited set of combinations of factor levels to be observed, from the large number of conceivable combinations. An example is the ecological simulation case-study with 281 parameters in Bettonvil and Kleijnen (1997); obviously the number of combinations is at least 2^{281} (which is a huge number, exceeding 10^{84}). An example with fewer factors (less than, say, fifteen) may be a supermarket simulation. In a simulation context, I define *DOE* as selecting the combinations of factor levels that will be actually simulated when experimenting with the simulation model. A popular type of design is the so-called 2^{k-p} design: k factors are changed in the experiment; each factor has two levels; only a fraction (namely 2^{-p} with $p = 0, 1, ...$) of the 2^k combinations is actually simulated. Depending on the

size of that fraction, the resolution of the design is 3, 4, 5 , ... : unbiased estimators of main effects only, sums of two-factor interactions, individual two-factor interactions, ...

After selecting the combinations of factor levels, the simulation program is executed or 'run'. Next the resulting I/O data of the simulation experiment are analyzed, applying *ANOVA* or *regression analysis*. This analysis estimates the importance of the individual factors (sensitivity analysis); that is, statistically significant factors may be considered to be important (the usual caveat about type I and type II errors applies; also see the next section, §3). In the simulation field such a regression model is called a *metamodel*, since it is a model of the I/O behavior of the underlying simulation model; see Friedman (1996), Kleijnen (1987). (Some call the metamodel a response surface, a repromodel, or a compact model.)

Typically, this metamodel uses one of the following three polynomial approximations.

(i) A first-order polynomial, which consists of an overall or grand mean β_0 and k main effects (say) β_j with $j = 1, \ldots, k$.

(ii) A first-order polynomial augmented with interactions between pairs of factors (two-factor interactions) $\beta_{j;\,j'}$ with $j' = j + 1, \ldots, k$.

(iii) A second-order polynomial, which adds purely quadratic effects $\beta_{j;\,j}$ to (ii).

Obviously, the first-degree polynomial in (i) misses interactions, and has constant marginal effects. Extending the second-order polynomial in (iii) to a third-order polynomial would be more difficult to interpret; it would also need many more simulation runs to estimate its many parameters β. So a second-order polynomial may be a good compromise, depending on the goal of the metamodel. Anyhow, an important practical question is: How should analysts select a particular degree for the polynomial approximation, and how should they validate the resulting metamodel?

To answer this question, some analysts use the well-known *multiple correlation coefficient* R^2. For example, Kleijnen (1995a) fits second-order polynomials, which give multiple correlation coefficients that - for the four scenarios simulated - range between 0.96 and 0.98 (also see below)

More refined selection procedures and tests use sequential DOE combined with cross-validation and Rao's F test; see Kleijnen and Sargent (1999) and Kleijnen, Cheng, and Feelders (1998).

A case study that does explicitly demonstrate the role of DOE and regression analysis in validation, is the ecological simulation in Bettonvil and Kleijnen (1997) and Kleijnen, Van Ham, and Rotmans (1992). The regression metamodel in the latter article helped to detect a serious error in the simulation model: one of the original modules should be split into two modules. Both publications further show that some factors are more important than the ecological experts originally expected; this 'surprise' gives additional insight into the simulation model.

Another case study is the sonar simulation in Kleijnen (1995a). This simulation model consists of several modules (submodels). There are no data for the modules 'inside' the model (these modules are not at the input or output boundary of the model). For each such module, a second-order polynomial is specified as metamodel. To estimate a second-order polynomial, Kleijnen (1995a) uses a *central composite design*, as analysts often do. This design combines a 2^{k-p} design with a one-factor-at-a-time design, plus one 'central' combination, which is at the center of the experimental area. For two modules the following results are found.

For one module, the naval experts suggest that its two factors have specific signs (namely $\beta_2 > 0$, $\beta_3 < 0$, $\beta_{2,3} < 0$). Indeed do the corresponding estimates turn out to have these signs. So this module has the correct I/O transformation, and its validity does not seem questionable. Of course, it cannot be claimed that its validity has been proven statistically!

The other module has six factors, and the central composite design has as many as 77 factor combinations. It turns out that one of these six factors has no significant effects at all: no main effect, no interactions with the other five factors, no quadratic effect. These results agree with the experts' qualitative knowledge. So the validity of this module is not questioned either.

These case studies illustrate that DOE with its regression analysis treats the simulation model as a *black box*: the simulation model's I/O is observed, and the factor effects in the metamodel are estimated. An advantage is that DOE can be applied to all simulation models, either deterministic or stochastic, discrete-event or continuous (a disadvantage is that DOE cannot exploit the specific structure of a given simulation model).

DOE assumes that the area of experimentation is given. A valid simulation model, however, requires that the inputs be restricted to a certain domain of factor combinations. This domain corresponds with the *experimental frame* in Zeigler (1976); also see Trybula (1994).

Related to sensitivity analysis is *risk analysis* or *uncertainty analysis*. Risk analysis also runs a simulation model for various combinations of factor levels. Risk analysis is performed because the input parameter values of the simulation model are not accurately

known; therefore risk analysis samples from a prespecified (joint) probability distribution for these parameters. This sampling uses the Monte Carlo technique (sometimes refined to Latin hypercube sampling or LHS; see Helton et al. 1997). So typically, its number of combinations is much larger than in sensitivity analysis using DOE.

I think that the basic difference between sensitivity analysis and risk analysis is that the latter tries to answer the question: what is the probability of a *disaster*? That disaster may be a nuclear accident, an ecological collapse, a financial mis-investment, etc. These disasters are *unique events*, whereas the case studies above concern repetitive events (e.g., average customer waiting time, mine detection probability). Consequently, validation in risk analysis is very difficult; see Jansen and De Vries (1998). A better term may be *credibility*; also see Fossett, Harrison, Weintrob, and Gass (1991) and Hodges (1991).

I would further add that from a risk analysis viewpoint, DOE selects extreme combinations of factor values that have very low probability of realization. Risk analysis, however, samples from the whole domain of possible combinations, according to the prespecified input distribution.

Risk analysts try to improve the underlying model's *credibility* by applying certain *statistical* techniques. For example, they apply regression analysis to detect which factors have significant effects; next - using their expert knowledge - they try to explain why these factors are important. An example is the following case study.

To obtain permission for nuclear waste disposal in the waste-isolation pilot-plant (WIPP) near Carlsbad, New Mexico (NM), a simulation model was developed at Sandia National Laboratories (SNL) in Albuquerque (NM). The Environmental Protection Agency (EPA) will give permission to start using the WIPP, only if the simulation model is accepted as credible - and the model's output shows an acceptable risk. Details on statistical techniques are given by Helton et al.(1997) and Kleijnen and Helton (1999).

3. REAL OUTPUT DATA: CLASSIC TESTS

How realistic is it to assume that there is data on the output - not the input - of the real system? Let us return to the case study on the search for mines by means of sonar, reported by Kleijnen (1995a). In this case study it is impossible to measure the environment - namely, the temperature and the salinity of the sea water that affect sonar performance - at all times and places. To obtain real output data on the detection of mines, the navy has one team deposit mines on the sea bottom; next another team searches for these mines (in general, the military conducts field tests; likewise, private companies build pilot plants to obtain data). In general, if the real-world scenarios are not measured, then only the outputs of the real and the simulated systems can be compared.

Note that in some situations the analysts are 'drown by the numbers'; examples are data on supermarket sales and on telecommunication operations. In general, data are abundant if systems are *electronically monitored*; examples are point of sale systems (POSS) and electronic data interchange (EDI). Another example is the milk robot simulation in Halachmi et al. (1999): cows are monitored electronically (also see the next section, §4)H0.

Let us return to the supermarket example. Suppose that the real output (say) x is the 90% quantile of the individual (autocorrelated) waiting times w_t of the customers served per day in the real system (the manager is assumed to be interested in 'excessive' waiting times, not in the mean waiting time; neither is she interested in the whole time path generated by the simulation run). Likewise, the simulated output (say) y is the 90% quantile of the individual (autocorrelated) waiting times v_t of the customers served per day in the simulated system. Suppose further that n days are observed in the real system, and m days are simulated. This yields $w_{i;\,t}$, waiting time of customer t on day i with $i = 1, ..., n$ in the real system. Analogously we have $v_{j;\,t}$ with $j = 1, ..., m$. This gives x_i, the 90% quantile of $w_{i;\,t}$ and y_j, the 90% quantile of $v_{j;\,t}$. Assume that each real or simulated day gives an independent and identically distributed (i.i.d.) observation (no seasonality; only busy Saturdays are measured).

The ideal simulation model would have a statistical distribution function for its output (say) F_y that is identical to the distribution for the real system F_x (also see Nayani and Mollaghasemi 1998, and Rao, Owen, and Goldsman 1998) In practice, however, the manager is not interested in the whole distribution F_x, but only in particular characteristics, the most popular being the mean, $E(x) = \mu_x$. For example, the 90% quantile varies from day to day, but its expected value is taken as the criterion to manage the supermarket. (In the next section we shall see how both the mean and the variance of x can be taken into account when validating a simulation model. However, if the purpose of the simulation is to help manage $E(x)$, then $var(x) = \sigma_x^2$ may be ignored.)

Define the mean difference $\mu_d = \mu_x - \mu_y$ Then the n and m observations on the real and the simulated

systems respectively give the classic estimators $\bar{x}$, $\bar{y}$, s_x^2, and s_y^2 of the means and variances of *x* and *y*. These estimators yield *two-sample Student's t statistic* with $n + m - 2$ degrees of freedom:

$$t_{n+m-2} = \frac{(\bar{x} - \bar{y}) - \mu_d}{[(n-1)s_x^2 + (m-1)s_y^2]^{1/2}} \cdot \frac{[(n+m-2)nm]^{1/2}}{(n+m)^{1/2}} \quad (1)$$

Obviously, the null-hypothesis is that simulated and real means are equal; that is, H_0: $\mu_d = 0$. The power of this test increases, as in Equation (1) $|\mu_d|$ increases (bigger differences are easier to detect), *n* or *m* increases (more days simulated or measured), or σ_x or σ_y decreases (less noise: more customers per day or lower traffic rate).

Note that defining $d = x - y$ means $\sigma_d^2 = \sigma_x^2 + \sigma_y^2 - 2\rho_{x;y}\sigma_x\sigma_y$ so the analysts may try to create a positive linear correlation between *x* and *y* - see $\rho_{x;y}$ - through the use of trace-driven simulation: see the next section (§4).

A type II error is likely to be committed if only a few days are simulated or there is much noise: an important difference (H_0: $|\mu_d| >> 0$) may go undetected (non-significant *t*). A type I error is also possible: if very many data are available, then an unimportant difference between the simulated and the real responses (H_0: $|\mu_d| = \epsilon$) can give a significant *t*-value.

Unfortunately, the test in Equation (1) assumes that the outputs *x* and *y* are normal (Gaussian) besides i.i.d., denoted as n.i.i.d.. Simulation models, however, may give non-normal outputs. The t statistic is known to be not very sensitive to nonnormality. Nevertheless, outputs such as estimated quantiles may show serious non-normality.

Let us briefly return to the sonar case study. This application gives a *binary* response variable: detect or miss a mine. The *m* simulation runs give a binomial variable with parameters *m* and (say) *p*, the detection probability. Analogously, the field test gives a binomial variable with parameters *n* and *q*. To test the null-hypothesis of equal simulated and real probabilities (H_0: $p = q$), Kleijnen (1995a) uses the *t*-statistic as an approximate test. Another case-study that applies this *t*-test is the traffic simulation by Rao et al. (1998).

An alternative to the *t* test is *Johnson's modified Student statistic*, which includes an estimator for the skewness of the output distribution; see Johnson (1978) and Kleijnen, Kloppenburg, and Meeuwsen (1986).

Another alternative is the class of *distribution-free* tests (such as the rank test); see Conover (1971). *Jackknifing* is also a robust technique, which requires only slightly more computer time for the analysis of the simulation output; see Efron and Tibshirani (1993). In practice, however, these alternatives are rarely applied - unfortunately. An application of a distribution-free (Kolmogorov-Smirnov) test is given by Rao et al. (1998).

One more alternative statistical technique is *bootstrapping*, which is a type of Monte Carlo simulation; see Efron and Tibshirani (1993). We shall return to bootstrapping, in the next section.

4. REAL I/O DATA: TRACE-DRIVEN

Comparing data on the real and the simulated systems makes more sense if both systems are observed under *similar scenarios*; for example, a busy day at the real supermarket should be compared with a busy day at the simulated store. More specifically, in queueing systems such a supermarkets input data consists of customers' arrival times and cashiers' service times, whereas output data concerns customers' waiting times. Trace-driven simulation means that the analysts feed real input data into the simulation program, in *historical order*. After running the simulation program, the analysts compare the time series of simulated output with the historical time series of real output. But how should they make this comparison? What is wrong with the following *naive* analysis of trace-driven simulation?

Make a scatter plot with (say) *x* and *y* - real and simulated outputs that use the same input. Fit a line $y = \beta_0 + \beta_1 x$, and test whether $\beta_1 = 1$ and $\beta_0 = 0$; see Figure 1 taken from the case study in Kozempel, Tomasula, and Craig (1995). (This validation procedure is also recommended by Van der Zouwen and Van Dijkum 1998.)

It is easy to prove that this analysis tends to reject a valid simulation model too often. Indeed, suppose the

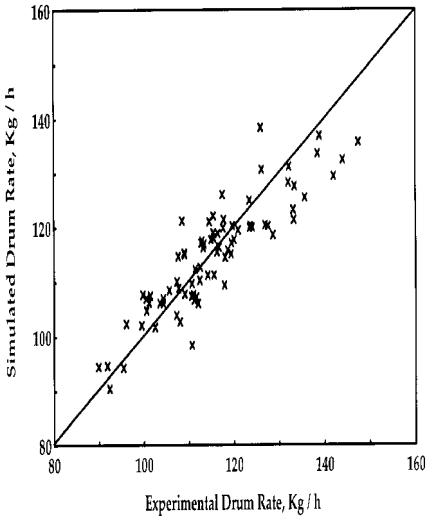

Figure 1: Example of Wrong Validation of Trace - driven Simulation (Source: Kozempel et al. 1995, p. 232)

simulation model is valid in the sense that the real and the simulated outputs have the same mean $\mu_x = \mu_y$ (= μ) and the same variance $\sigma_x^2 = \sigma_y^2$ (= σ^2). Suppose further that this mean is positive ($\mu > 0$) - as is the usual case in queueing simulations - and that the simulation model is not perfect ($\rho_{xy} < 1$). In general, for the linear regression model $y = \beta_0 + \beta_1 x$ we have $\beta_1 = \rho_{xy} \sigma_y / \sigma_x$ and $\beta_0 = \mu_y - \beta_1 \mu_x$. Hence, a valid simulation model gives $0 < \beta_1 < 1$ and $0 < \beta_0 < \mu$. So if the analysts test whether $\beta_1 = 1$ and $\beta_0 = 0$, then they are likely to reject the valid simulation model!

This is indeed what happens in Lysyk (1989): he finds an estimated slope significantly smaller than unity and an intercept significantly positive. Since he expects a unit slope and a zero intercept, he tries to explain this phenomenon away. Figure 1 also suggests $\beta_0 > 0$ and $\beta_1 < 1$ (we cannot give the actual estimates since we do not have the original numbers in Figure 1 available). More examples will follow below.

A *novel* validation test for trace-driven simulation is derived by Kleijnen, Bettonvil, and Van Groenendaal (1996, 1998). They compute not only the n differences d_i (also see Equation 1 with $n = m$), but also the n sums (say) $q_i = x_i + y_i$. Next they fit a line $d = \gamma_0 + \gamma_1 q$ to these n pairs (d_i, q_i). Then they formulate the null-hypothesis H_0: $\gamma_0 = 0$ and $\gamma_1 = 0$. Obviously, this (joint, composite) hypothesis implies $\mu_d = 0$ or $\mu_x = \mu_y$. Moreover, assuming normality for x and y, it is easy to prove that $\gamma_1 = 0$ implies equal variances: $\sigma_x^2 = \sigma_y^2$. To test this joint hypothesis, standard regression software (which applies an F test) can be used.

Kleijnen et al. (1998) apply both the naive and the novel regression analyses to single server systems with Poisson arrival and service times (Markov systems with one server: M/M/1). This gives the following conclusions.

(i) The naive test rejects a truly valid simulation model substantially more often than the novel test does.

(ii) The naive test shows 'perverse' behavior in a certain domain; that is, the worse the simulation model is (in that domain), the higher is its probability of acceptance.

(iii) The novel test does not reject a valid simulation model too often (that is, it rejects with probability α), provided the outputs are transformed logarithmically to realize normality

Besides this academic M/M/1 study, there is a case study that applies both the naive and the novel regression analysis, namely the milk robot simulation in Halachmi et al. (1999). Again, the naive test rejects the simulation model much more often than the novel test does. Obviously, it is unknown whether this simulation model is valid or not: it is a real case study - unlike the academic study by Kleijnen et al. (1996, 1998).

Both the naive and the novel analyses assume n.i.i.d. (real and simulated) outputs. Kleijnen, Cheng, and Bettonvil (1999), however, consider the validation of simulation models with *non-normal* outputs. They study several test statistics, using bootstrapping. They conclude that actually the simplest test is best: bootstrapping the difference between the average simulated and real responses gives the correct type I error probability and has good power. They discuss in detail how to bootstrap the real and the simulated outputs.

5. CONCLUSIONS

In practice, validation has many forms, but I focused on validation through mathematical statistics. Statistical validation may use various tests, depending on the type of data available for the real system. I discussed the following three situations..

(i) *No* real data

Even if there is no data on the input or output of the real system, the analysts can still generate *simulated* data. More specifically, the analysts should perform sensitivity analysis to find out whether the simulation model contradicts *qualitative, expert knowledge*. If the simulation‘s input/output (I/O) behavior violates this knowledge, the model should be seriously searched for programming and conceptual errors. This sensitivity analysis should be guided by DOE including regression metamodels; an inferior approach changes only one factor at a time.

(ii) Only data on real *output*

If there is data on the output of the real system, the means of real and simulated output distributions may be compared through the two-sample Student *t* test. Alternatives are Johnson's modified *t* statistic (estimating the skewness of the output distribution), distribution-free statistics, and bootstrapping.

(iii) *I/O* data on real system

Real input data enable trace-driven simulation. The validation of this type of simulation, however, should not use a scatter plot with real and simulated outputs, testing whether the fitted line has unit slope and zero intercept. Instead, two alternatives were discussed. Alternative #1 regresses differences on sums; this analysis applies if the outputs are n.i.i.d. Alternative #2 uses bootstrapping of a simple validation statistic based on differences; this provides acceptable type I and II errors.

To demonstrate the applicability of the various statistical methods, I summarized several case studies. Nevertheless, because validation involves the art of modeling and the philosophy of science, validation will remain controversial!

REFERENCES

Beck, M.B., J.R. Ravetz, L.A. Mulkey, and T.O. Barnwell (1997), On the problem of model validation for predictive exposure assessments. *Stochastic Hydrology and Hydraulics*, 11, pp. 229-254

Bettonvil, B. and J.P.C. Kleijnen (1997) Searching for important factors in simulation models with many factors: sequential bifurcation. *European Journal of Operational Research*, 96, no. 1, pp. 180-194

Conover, W.J. (1971), *Practical non-parametric statistics*. Wiley, New York

Efron, B. and R.J. Tibshirani (1993), *Introduction to the bootstrap*. Chapman & Hall, London

Fossett, C.A., Harrison D., Weintrob H., and Gass S.I. (1991), An assessment procedure for simulation models: a case study, *Operations Research* 39, pp. 710-723

Friedman, L.W. (1996), *The simulation metamodel*. Kluwer, Dordrecht, Netherlands

Halachmi, I et al. (1999), Validation of a simulation model in robotic milking barn design. Working paper, Institute of Agricultural and Environmental Engineering (IMAG-DLO), Wageningen, Netherlands

Helton, J.C., D.R. Anderson, M.G. Marietta, and R.P. Rechard (1997), Performance assessment for the waste isolation pilot plant: from regulation to calculation for 40 CFR 191.13. *Operations Research*, 45, no. 2, pp. 157-177

Hodges, J.S. (1991), Six (or so) things you can do with a bad model. *Operations Research*, 39, no. 3, pp. 355-365

Jansen, M. and B. De Vries (1998), Global modelling: managing uncertainty, complexity and incomplete information. *Validation of simulation models*, eds. C. van Dijkum, D. de Tombe, and E. van Kuijk, SISWO, Amsterdam

Johnson N.J. (1978), Modified t tests and confidence intervals for asymmetric populations. *Journal of the American Statistical Association*, 73, pp. 536-544

Kleijnen, J.P.C. (1999), Statistical validation of simulation, including case studies. *Validation of simulation models*, eds. C. van Dijkum, D. de Tombe, and E. van Kuijk, SISWO, Amsterdam

--- (1998), Experimental design for sensitivity analysis, optimization, and validation of simulation models. *Handbook of simulation*, ed. J. Banks, Wiley, New York

--- (1995a), Case study: statistical validation of simulation models. *European Journal of Operational Research*, 87, no. 1, pp. 21-34

--- (1995b), Verification and validation of simulation models. *European Journal of Operational Research*, 82, no. 1, pp. 145-16

--- (1987) *Statistical tools for simulation practitioners*. Marcel Dekker, New York

---, B. Bettonvil, and W. Van Groenendaal (1998). Validation of trace-driven simulation models: a novel regression test. *Management Science*, 44, no. 6, pp. 812-819

---, B. Bettonvil, and W. Van Groenendaal (1996). Validation of trace-driven simulation models: regression analysis revisited. *Proceedings of the 1996 Winter Simulation Conference*, eds. J.M. Charnes, D.J. Morrice, D.T. Brunnner, and J.J. Swain, pp. 352-359

---, R.C.H. Cheng, and B. Bettonvil (1999),Validation of trace-driven simulation models: bootstrapped tests. Working Paper

---, R.C.H. Cheng, and A.J. Feelders (1998), Bootstrapping and validation of metamodels in simula-

tion. *Proceedings of the 1998 Winter Simulation Conference*, eds. D.J. Medeiros, E.F. Watson, J.S. Carson, M.S. Manivannan, pp. 701-706

--- and J. Helton (1999),Statistical analysis of scatter plots to identify important factors in large-scale simulations. *Reliability Engineering and Systems Safety*, 65, no. 2, pp. 147-1197

---, G.L.J. Kloppenburg, and F.L. Meeuwsen (1986), Testing the mean of an asymmetric population: Johnson's modified t test revisited. *Communications in Statistics, Simulation and Computation*, 15, no. 3, pp. 715-732

--- and Ő. Pala (1999), Maximizing the simulation output: a competition. *Simulation* (accepted)

--- and R.G. Sargent (1999), A methodology for the fitting and validation of metamodels in simulation. *European Journal of Operational Research* (in press)

---, G. Van Ham, and J. Rotmans (1992), Techniques for sensitivity analysis of simulation models: a case study of the CO2 greenhouse effect. *Simulation*, 58, no. 6, pp. 410-417

Knepell, P.L. and D.C. Arangno (1993), Simulation validation; a confidence assessment methodology. IEEE Computer Society Press, Los Alamitos

Kozempel, M.F., P. Tomasula, and J.C. Craig (1995), The development of the ERRC food process simulator, *Simulation; Practice and Theory*, 2, 4-5

Law A.M., and W.D. Kelton (1991), *Simulation modeling and analysis; Second Edition*, McGraw-Hill, New York

Lysyk, T.J. (1989), Stochastic model of Eastern spruce budworm (lepidoptera: tortricidae) phenology on white spruce and balsam fir, *Journal of Economic Entomology*, 82, 4, 1161-1168

Nayani, N. and M. Mollaghasemi (1998), Validation and verification of the simulation model of a photolithography process in semiconductor manufacturing. *Proceedings of the 1998 Winter Simulation Conference*, eds. D.J. Medeiros, E.F. Watson, J.S. Carson, and M.S. Manivannan, pp. 1017-1022

Rao, L., L. Owen, and D. Goldsman (1998), Development and application of a validation framework for traffic simulation models. *Proceedings of the 1998 Winter Simulation Conference*, eds. D.J. Medeiros, E.F. Watson, J.S. Carson, and M.S. Manivannan, pp. 1079-1086

Sargent, R.G. (1996), Verifying and validating simulation models. *Proceedings of the 1996 Winter Simulation Conference*, eds. J.M. Charnes, D.M. Morrice, D.T. Brunner, and J.J. Swain, pp. 55-64

Trybula, W.J. (1994), Building simulation models without data. *Proceedings of the International Conference on Systems, Man, and Cybernetics*, IEEE, pp. 209-214

Van der Zouwen, J. and C. van Dijkum (1998), Towards a methodology for the empirical testing of complex social cybernetic models. XIVth ISA World Congress of Sociology

Zeigler, B. (1976) *Theory of modelling and simulation*. Wiley Interscience, New York

AUTHOR BIOGRAPHY

JACK P.C. KLEIJNEN is a Professor of Simulation and Information Systems. His research concerns simulation, mathematical statistics, information systems, and logistics; this research resulted in six books and nearly 160 articles. He has been a consultant for several organizations in the USA and Europe, and has served on many international editorial boards and scientific committees. He spent several years in the USA, at both universities and companies, and received a number of international fellowships and awards. More information is provided on his web page: http://center.kub.nl/staff/kleijnen